
Proceedings Paper:
Dufresne, Emilie Sonia orcid.org/0000-0001-9290-7037, Edwards, Parker B., Harrington, Heather A et al. (1 more author) (2020) Sampling real algebraic varieties for topological data analysis. In: 18th IEEE International Conference on Machine Learning and Applications:ICMLA 2019. IEEE International Conference On Machine Learning And Applications, 16-19 Dec 2019 IEEE , USA .

https://doi.org/10.1109/ICMLA.2019.00253

Sampling real algebraic varieties for topological data analysis

Emilie Dufresne
Department of Mathematics
University of York
York, UK
emilie.dufresne@york.ac.uk

Parker B. Edwards
Department of Mathematics
University of Florida
Gainesville, FL, USA
pedwards@ufl.edu

Heather A. Harrington
Mathematical Institute
University of Oxford
Oxford, UK
harrington@maths.ox.ac.uk

Jonathan D. Hauenstein
Department of Applied and Computational Mathematics and Statistics
University of Notre Dame
Notre Dame, IN, USA
hauenstein@nd.edu

***Abstract*—Topological data analysis (TDA) provides tools for computing geometric and topological information about spaces from a finite sample of points. We present an adaptive algorithm for finding provably dense samples of points on real algebraic varieties given a set of defining polynomials for use as input to TDA. The algorithm utilizes methods from numerical algebraic geometry to give formal guarantees about the density of the sampling, and also employs geometric heuristics to reduce the size of the sample. As TDA methods consume significant computational resources that scale poorly in the number of sample points, our sampling minimization makes applying TDA methods more feasible. We provide a software package that implements the algorithm, and showcase it through several examples.**

***Index Terms*—topological data analysis, real algebraic varieties, dense samples, numerical algebraic geometry, minimal distance**

I. Introduction

Understanding the geometry and topology of real algebraic varieties is an ubiquitous and challenging problem in applications modelled by polynomial systems. For kinematics problems, geometric insight about configuration spaces can lead to physical insights (e.g., [41]), while the geometry of varieties provides information about biochemical systems (e.g., [33]). Here, we present a new algorithm fulfilling a key step in applying topological data analysis methods (TDA), particularly persistent homology [56] (PH), to real algebraic varieties. The algorithm takes as input a list of polynomials defining a real algebraic variety, and outputs a sample of points on the variety tailored for input to PH.

A. Prior work

PH computes topological features closely related to a variety's Betti numbers. Several analyses and proposed algorithms (e.g., [4], [22], [52]) offer theoretical complexity guarantees for variants of the problem of computing Betti numbers given a list of defining polynomials as input; however, implementations are not available. Other approaches similar to PH take as input a sample of points from a variety, with output that can be used to estimate Betti numbers. Extensive effort has produced a large number of surface reconstruction algorithms, particularly for nonsingular surfaces embedded in $\mathbb{R}^3$ (e.g., [2], [10], [24], [30], [42]). Unlike PH, none of these methods apply to general real varieties. There is a probabilistic algorithm for computing Betti numbers from uniform random point samples [48]. Given a sample of points from a real variety, one may alternatively compute other features. For a large enough sample of "general" points, [47] studies the "Betti diagram" of a projective variety. Given set of general points, one can "learn" the equations defining the algebraic variety [13].

The algorithm in [12] produces samples from the uniform distribution on a variety. Among deterministic sampling approaches, subdivision and reduction sampling methods [46], [54] most closely resemble our algorithm. These methods can take the polynomials defining a real semialgebraic set as input and output a dense sample of points. For PH computations, they exhibit two drawbacks: (1) Sample points in the output need not be especially close to the underlying variety (δ is not small in the sense of Definition II.8). (2) Adjusting current implementations to reduce the number of sample points is not straightforward.

Our approach for sampling varieties is based on numerical algebraic geometry, with the books [5], [55] providing a general overview. The algorithm addresses the first point above by constructing provably dense samples with points very close to the underlying variety. The theoretical version of the algorithm can be readily adjusted to incorporate geometric heuristics which significantly reduce the number of points in the final output, thereby addressing the second point. An implementation is publicly available as the Python package `tdasampling` on PyPI and the package source code is available at https://github.com/P-Edwards/tdasampling.

B. Organization

The paper is organized as follows: we recall TDA theory and computations in Section II, and numerical algebraic geometry in Section III. Section IV details the sampling algorithm, proves its correctness, and discusses the geometric heuristics for sample minimization. In Section V, we illustrate our sampling algorithm with TDA on several examples.

II. Topological data analysis

Topological data analysis is a field of research encompassing theory and algorithms which adapt the theory of topology and

geometry to analyze the "shape" of data. The goal of our sampling algorithm is to produce input for TDA algorithms.

We apply the persistent homology pipeline popularized by Carlsson in [16], and summarized by Ghrist in [32]. Broader overviews of other TDA methods can be found in [19], [27], [49], [50]. The PH pipeline takes as input a *point cloud* of finitely points in $\mathbb{R}^N$, or a distance matrix. It computes and outputs an summary of the algebraic topological features of the input. See e.g. [27], [50] for detailed discussions of PH, and [35] for an introduction to homology. All homology discussed is with coefficients in a field.

A. Building simplicial complexes from data

Definition II.1. Let $\hat{X}$ be a finite subset of a metric space Y, and $\epsilon \geqslant 0$ be a real number. The *Čech complex* for $\hat{X}$ with parameter ϵ, $C_\epsilon(\hat{X})$, is the nerve of the set $\{\bar{B}_x(\epsilon)\}_{x\in\hat{X}}$ where $\bar{B}_x(\epsilon)$ denotes the closed ball of radius ϵ with center x, and the *Vietoris-Rips complex*, $R_\epsilon(\hat{X})$, is the flag complex of $C_{\frac{\epsilon}{2}}(\hat{X})$. See e.g. [27, §3.2] for details.

Vietoris-Rips complexes are typically cheaper to compute than Čech complexes, again see e.g. [27, §3.2]. The following interleaving result precisely describes a manner in which Vietoris-Rips complexes estimate Čech complexes.

Theorem II.2 (de Silva and Ghrist [23])**.** *If $\hat{X}$ is a finite set of points in $\mathbb{R}^N$ and $\epsilon > 0$ there is a chain of inclusions*

$$C_{\frac{\epsilon'}{2}}(\hat{X}) \subseteq R_{\epsilon'}(\hat{X}) \subseteq C_\epsilon(\hat{X}) \subseteq R_{2\epsilon}(\hat{X})$$

whenever $\frac{\epsilon}{\epsilon'} \geqslant \frac{1}{2}\sqrt{\frac{2N}{N+1}}$.

B. Persistent homology

We summarize the categorical approach to PH introduced in [15].

Definition II.3. Let k be a field ($\frac{\mathbb{Z}}{2\mathbb{Z}}$ in all subsequent examples). A *persistence module* is a functor $F : (\mathbb{R}, \leqslant) \to \mathbf{vect}_k$ from the poset $(\mathbb{R}, \leqslant)$ to the category $\mathbf{vect}_k$ consisting of (finite dimensional) vector spaces over k with linear maps between them. Explicitly, F is determined by:

- A k-vector space $F(\epsilon)$ for every $\epsilon \in \mathbb{R}$
- A linear map $F(\epsilon \leqslant \epsilon') : F(\epsilon) \to F(\epsilon')$ for every pair of real numbers $\epsilon \leqslant \epsilon'$ such that:
  - $F(\epsilon \leqslant \epsilon)$ is the identity map from $F(\epsilon)$ to itself
  - Given real numbers $\epsilon \leqslant \epsilon' \leqslant \epsilon''$, $F(\epsilon \leqslant \epsilon'') = F(\epsilon' \leqslant \epsilon'') \circ F(\epsilon \leqslant \epsilon')$

Definition II.4. A point $\epsilon \in \mathbb{R}$ is *regular* for a persistence module F if there exists an open interval $I \subseteq \mathbb{R}$ such that $\epsilon \in I$ and $F(a \leqslant b)$ is an isomorphism for all pairs $a \leqslant b \in I$. Otherwise ϵ is *critical*. A functor is *tame* if it has finitely many critical values.

Example II.5. For any finite point cloud $\hat{X} \subseteq \mathbb{R}^N$ and real numbers $0 \leqslant \epsilon \leqslant \epsilon'$, there is an inclusion $C_\epsilon(\hat{X}) \subseteq C_{\epsilon'}(\hat{X})$. Fixing $p \geqslant 0$ and applying H_p results in a sequence of vector spaces and $\frac{\mathbb{Z}}{2\mathbb{Z}}$-linear maps $H_p(C_\epsilon(\hat{X})) \to H_p(C_{\epsilon'}(\hat{X}))$ induced by inclusion. The assignment $\epsilon \mapsto H_p(C_\epsilon(\hat{X}))$ along with these linear maps defines a tame persistence module $HC := H_pC_\bullet(\hat{X})$. An analagous persistence module exists for the Vietoris-Rips complex denoted by HR. ◁

Definition II.6. The *rank function* of a tame module F assigns $x \leqslant y \mapsto \operatorname{rank} F(x \leqslant y)$ for every $x \leqslant y \in \mathbb{R}$. Let $\bar{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$. The *persistence diagram* of F is the multiset DF of points $(b, d) \in \bar{\mathbb{R}}^2$ uniquely determined by the following two conditions: (i) $b \leqslant d \in \bar{\mathbb{R}}^2$, and (ii) for $x \leqslant y \in \mathbb{R}$, $\operatorname{rank}(x \leqslant y)$ is the number of points in DF above and to the left of (x, y).

Theorem II.7 (Fundamental Theorem of Persistent Homology)**.** *Let F and G be tame persistence modules. F and G are isomorphic if and only if "decorated" versions (see [50] §1.3) of DF and DG are equal.*

The original algebraic version of Theorem II.7 for PH appears in [56], and a categorical version in [15]. Each point (x, y) in a module's persistence diagram can be viewed as describing the range of parameter values through which a single independent feature in the module persists. See Fig. 1.

C. Computational considerations

Persistence diagrams for modules arising from the homology of finite simplicial complexes can be computed via the Persistence Algorithm (see e.g. [27] ṾII.1). In practice, memory consumption that grows rapidly with the number of input sample points is the limiting factor in computations. See [49] for details on computational costs, and e.g. [8], [20], [43] for various optimization strategies.

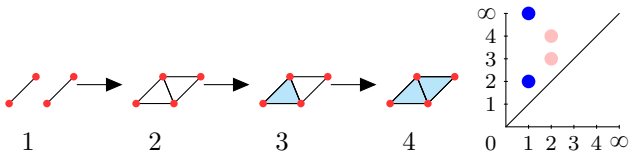

Fig. 1: The left figure shows a filtered complex as it changes with parameter value. The right figure depicts the persistence diagram. Blue points represent 0 dimensional homology and pink points represent 1 dimensional homology.

D. Homology inference

Recall that any compact topological space $Y \subseteq \mathbb{R}^N$ defines the distance-to-Y function $d_Y : \mathbb{R}^N \to \mathbb{R}$. The function is given by $d_Y(z) = \min_{y\in Y} d(z, y)$ for any $z \in \mathbb{R}^N$. Given any real number $\epsilon \geqslant 0$, define $Y^\epsilon = d_Y^{-1}(-\infty, \epsilon]$. The space Y^ϵ is formed from Y by taking the union of all closed balls of radius ϵ in $\mathbb{R}^N$ centered at points of Y.

Definition II.8. Let $A, B \subseteq \mathbb{R}^N$ be compact and $0 \leqslant \delta \leqslant \epsilon \in \mathbb{R}$. The set A is a (δ, ϵ)-*sample* of B if $A \subseteq B^\delta$ and $B \subseteq A^\epsilon$.

Remark II.9. Definition II.8 is a specific instance of an *interleaving* between generalized persistence modules as defined in [14]. It is also generalization of the Hausdorff distance between subsets of metric space.

Definition II.10. Let $X \subseteq \mathbb{R}^N$ be a compact metric space. The homological feature size of X, $\mathrm{hfs}(X)$, is the infimum of all positive critical values over all dimensions p of the persistence module $\epsilon \mapsto H_p(X^\epsilon)$ (negative ϵ are assigned $\varnothing$).

Remark II.11. The homological feature size of a space X was introduced in [21], and is bounded below by the space's *reach* [1] and the space's *weak feature size* [17]. More precisely, $0 \leqslant \mathrm{reach}(X) \leqslant \mathrm{wfs}(X) \leqslant \mathrm{hfs}(X)$. The weak feature size of real semialgebraic sets is known to be positive ([31] §5.3), and so the homological feature size is positive as well. Compact real semialgebraic sets are absolute neighborhood retracts (see e.g. Theorem 3 of [40] and Corollary 3.5 of [34]), and in particular this implies $H_p(X)$ is isomorphic to $H_p(X^\kappa)$ for any compact semialgebraic set, $0 \leqslant \kappa < \mathrm{hfs}(X)$, and $p \geqslant 0$.

Theorem II.12 (Homology Inference Theorem, [18], [21])**.** *Let $\hat{X}, X \subseteq \mathbb{R}^N$, with X compact semialgebraic and $\hat{X}$ a finite (δ,ϵ)-sample of X, where $0 \leqslant \delta \leqslant \epsilon$ and $\mathrm{hfs}(X) > 2(\epsilon+\delta)$. Letting $HC = H_pC_\bullet(\hat{X})$, the dimension of $H_p(X)$ is the number of points in $D(HC)$ above and to the left of the point $(\epsilon, 2\epsilon+\delta) \in \bar{\mathbb{R}}^2$.*

Proof. From the definition of (δ,ϵ)-sample we have inclusions $X \hookrightarrow \hat{X}^\epsilon \hookrightarrow X^{\epsilon+\delta} \hookrightarrow \hat{X}^{2\epsilon+\delta} \hookrightarrow X^{2(\epsilon+\delta)}$. The Nerve Theorem (e.g. §4G.3 [35]) implies that $HC(a) \cong H_p(\hat{X}^a)$ for all $a \in \mathbb{R}$. Applying homology to the sequence and using the assumption on the homology when thickening X, we obtain the commutative diagram

$$H_p(X) \longrightarrow HC(\epsilon) \underbrace{\longrightarrow H_p(X) \longrightarrow HC(2\epsilon+\delta)}_{h} \longrightarrow H_p(X)$$

where the maps from $H_p(X)$ to itself are isomorphisms. Since there is an isomorphism from $H_p(X)$ to itself which factors through h, $\dim H_p(X) \leqslant \mathrm{rank}(h)$. The map h also factors through a map with domain $H_p(X)$, so $\mathrm{rank}(h) \leqslant \dim H_p(X)$. □

Corollary II.13. *Let $HC, X, \hat{X}, \epsilon$, and δ be as in Theorem II.12. The number of points above and to the left of $\left(2\epsilon\sqrt{\frac{N+1}{2N}}, 4\epsilon+2\delta\right)$ in the persistence diagram for $HR = H_pR_\bullet(\hat{X})$ is a lower bound for $\dim H_p(X)$.*

Proof. Let $a = 2\epsilon\sqrt{\frac{N+1}{2N}}$. By Theorem II.2, we have the following commutative diagram of linear maps

$$HR(a) \underbrace{\longrightarrow HC(\epsilon) \longrightarrow HC(2\epsilon+\delta) \longrightarrow}_{h} HR(4\epsilon+2\delta).$$

It follows that $\mathrm{rank}(h) \leqslant \mathrm{rank}(HC(\epsilon \leqslant 2\epsilon+\delta))$. Theorem II.12 (with $\epsilon' = a$) shows that the rank of $HC(\epsilon \leqslant 2\epsilon+\delta)$ is $\dim H_p(X)$. □

III. Sampling using numerical algebraic geometry

An algebraic variety $V \subseteq \mathbb{C}^N$ is the solution set of a system of polynomial equations. The set of real points of V, $V_\mathbb{R} = V \cap \mathbb{R}^N \subseteq \mathbb{R}^N$, is a real algebraic variety. One approach to compute a point on $V_\mathbb{R}$ is by computing a point $x \in V_\mathbb{R}$ which is a global minimizer of the distance function between a given test point $y \in \mathbb{R}^N$ and $V_\mathbb{R}$ [53]. We summarize the use of numerical algebraic geometry to perform this computation based on [37] (see also [3], [25], [51]), with Section IV relying on this to generate a provably dense sampling of $V_\mathbb{R}$.

Suppose that $V \subseteq \mathbb{C}^N$ is an algebraic variety of dimension d and $f(x)$ consists of $N-d$ polynomials such that V is the solution set of $f = 0$. This assumption simplifies formulating the critical point conditions of the minimization problem below, but can be relaxed, e.g., see [37]. Given a test point $y \in \mathbb{R}^N$, the approach of Seidenberg [53] is to compute a global minimizer of

$$\min \left\{ \textstyle\sum_{i=1}^N (x_i - y_i)^2 \;\middle|\; x \in V_\mathbb{R} \right\} \tag{III.1}$$

which is accomplished by solving the Fritz John optimality conditions, namely solving

$$G_y(x,\lambda) = \begin{bmatrix} f(x) \\ \lambda_0(x-y) + \sum_{i=1}^{N-d} \lambda_i \nabla f_i(x) \end{bmatrix}$$

on $\mathbb{C}^N \times \mathbb{P}^{N-d}$, where $\nabla f_i(x)$ is the gradient of $f_i(x)$ with respect to x and $\mathbb{P}^{N-d}$ is the $(N-d)$-dimensional projective space. Consider the homotopy

$$H_{y,\beta}(x,\lambda,t) = \begin{bmatrix} f(x) - t\beta \\ \lambda_0(x-y) + \sum_{i=1}^{N-d} \lambda_i \nabla f_i(x) \end{bmatrix}. \tag{III.2}$$

The following is immediate from coefficient-parameter continuation [45] showing that generic choices of parameter values (y,β) leads to a well-constructed homotopy $H_{y,\beta}$.

Proposition III.3. *There exists a nonempty Zariski dense open subset $U \subseteq \mathbb{C}^N \times \mathbb{C}^{N-d}$ such that if $(y,\beta) \in U$, then*

1) *the set $S \subseteq \mathbb{C}^N \times \mathbb{P}^{N-d}$ consisting of all solutions to $H_{y,\beta}(x,\lambda,1) = 0$ is finite and each is nonsingular;*
2) *the number of points in S is equal to the maximum number, as $y' \in \mathbb{C}^N$ and $\beta' \in \mathbb{C}^{N-d}$ both vary, of isolated solutions of $H_{y',\beta'}(x,\lambda,1) = 0$;*
3) *the solution paths defined by the homotopy $H_{y,\beta}(x,\lambda,t) = 0$ starting at the points in S at $t = 1$ are smooth for $t \in (0,1]$.*

The number of points in S is equal to the Euclidean distance degree of $\mathcal{V}(f-\beta)$ [25]. The set S can be computed using standard homotopy continuation as described in [37]. Since $G_y(x,\lambda) = H_{y,\beta}(x,\lambda,0)$, the endpoints E of the solution paths defined by $H_{y,\beta}(x,\lambda,t) = 0$ contained in $\mathbb{C}^N \times \mathbb{P}^{N-d}$ are solutions of $G_y = 0$. Hence, E is a finite set of solutions to $G_y = 0$ containing a global minimizer of (III.1) as stated in the following from [37, Thm. 5] and [51, Lemma 3.7].

Theorem III.4. *Let $(y,\beta) \in U$ where U is defined in Proposition III.3. Let E be the set of endpoints in $\mathbb{C}^N \times \mathbb{P}^{N-d}$ of the homotopy paths defined by $H_{y,\beta}(x,\lambda,t) = 0$. Define $\pi_1(x,\lambda) = x$. Then, $\pi_1(E) \cap V_\mathbb{R}$ contains finitely many points, one of which is a global minimizer of* (III.1)*. Hence, $V_\mathbb{R} = \varnothing$ if and only if $\pi_1(E) \cap V_\mathbb{R} = \varnothing$.*

Since $\pi_1(E) \cap V_\mathbb{R}$ consists of finitely many points, a global minimizer of (III.1) is identified by simply minimizing over these finitely many points.

IV. GENERATING SAMPLES

This section presents an algorithm integrating Theorem III.4 with geometric tools to produce provably dense samples of real algebraic varieties. The input and output are as follows.

Input:

- Polynomial system $f \subseteq \mathbb{R}[x_1, \ldots, x_N]$ defining a pure d-dimensional real algebraic variety $X = \mathrm{V}_{\mathbb{R}}(f)$.
- A compact region $R \subseteq \mathbb{R}^N$ which is a "box," i.e., of the form $R = [a_1, b_1] \times \cdots \times [a_N, b_N]$.
- A sampling density $\epsilon > 0$.
- An estimation error δ with $0 < \delta \leqslant \epsilon$.

Output: A (finite) set of points $\hat{X} \subset \mathbb{R}^N$ forming a (δ, ϵ)-sample of $X \cap R$.

Theorem III.4 provides a computationally tractable approach to finding very accurate estimated solutions of the optimization problem (III.1) for generic $y \in \mathbb{R}^N$. Following the terminology of Section III, we define the subroutine `MinDistance`. Theorem III.4 shows its output has the indicated properties.

Input : Polynomial system $f \subset \mathbb{R}[x_1, \ldots, x_N]$ defining a real variety $X := V_{\mathbb{R}}(f)$
Input : A (generic) point $y \in \mathbb{R}^N$
Input : Estimation error $\delta > 0$
Output: A set S of points where $s \in S$ has $d_X(s) \leqslant \delta$ for all $s \in S$ and $d_X(y) \leqslant \min_{s \in S} d_X(s) + \delta$.

1 $\beta \leftarrow$ a (generic) uniform random point in $\mathbb{C}^{N-d}$;
2 Solve the parameter homotopy $H_{y,\beta}$ in Eq. (III.2) using homotopy continuation (see e.g. [5]), returning set S;
3 `Return` (S)

Algorithm IV.1: MINDISTANCE(f, y, δ)

Input : A box $C = [c_1, d_1] \times \cdots \times [c_N, d_N]$
Output: Two boxes $R_1, R_2 \subseteq C$ with $C = R_1 \cup R_2$.

1 $j \leftarrow \arg\max_{i=1,\ldots,N} |d_i - c_i|$;
2 $R_1 \leftarrow [c_1, d_1] \times \cdots \times [c_j, \frac{c_j + d_j}{2}] \times \cdots \times [c_N, d_N]$;
3 $R_2 \leftarrow [c_1, d_1] \times \cdots \times [\frac{c_j + d_j}{2}, d_j] \times \cdots \times [c_N, d_N]$;
4 `Return` $(R_1$ *and* $R_2)$

Algorithm IV.2: SPLITBOX(C)

Definition IV.3. For any box $A \subseteq \mathbb{R}^N$, let T_A be the tree with root A whose nodes are boxes in $\mathbb{R}^N$. The children of any box C in T_A are the elements of `SplitBox`(C). The elements of `SplitBox`(C) have parent node C.

Remark IV.4. The key properties of `SplitBox` are that it breaks boxes into proper sub-boxes, that union of the sub-boxes is the original box, and that for any $\gamma > 0$ and box A, there is some n where all n-children of A in T_A have maximum side length at most γ.

Theorem IV.6. *Algorithm IV.5 terminates and outputs a (δ, ϵ)-sample of $X \cap R$.*

Lemma IV.7. *With notation as in Algorithm IV.5 (1) If M is a box in T_R with max side length at most $\frac{\epsilon - \delta}{\sqrt{N}}$, M will be marked "done" by Algorithm IV.5. (2) R is the union of regions marked "stop" by Algorithm IV.5.*

Input : Polynomial system $f \subseteq \mathbb{R}[x_1, \ldots, x_N]$, a box $R = [a_1, b_1] \times \cdots \times [a_N, b_N]$, sampling density $\epsilon > 0$, and estimation error $0 \leqslant \delta \leqslant \epsilon$
Output: A list of points which form a (δ, ϵ)-sample of $V_{\mathbb{R}}(f) \cap R$

1 Initialize an empty spatial database CoveredRegions which can store subregions of $\mathbb{R}^N$;
2 Initialize an empty list SampleOutput of points in $\mathbb{R}^N$;
3 **for** *each node M in T_R not marked "done", iterated via breadth first search* **do**
4 **if** *The maximum side length of M is at most $\frac{\epsilon - \delta}{\sqrt{N}}$ or M does not intersect any region stored in* CoveredRegions **then**
5 Run `MinDistance`(f, y, δ) where y is the center point of M, returning a set of sample points S with minimum distance D_y from y to any point in S;
6 Add regions $B_{D_y - \delta}(y)$ and $B_\epsilon(s)$ for each $s \in S$ to CoveredRegions. Add each $s \in S$ to SampleOutput.;
7 **end**
8 **if** *$M \subseteq B$ for any region B contained in* CoveredRegions **then**
9 Mark M with "stop", mark all nodes in the subtree rooted at M (including M) "done", and stop searching the subtree rooted at M.;
10 **end**
11 **if** *All unsearched boxes in T_R are marked "done"* **then**
12 End for loop.;
13 **end**
14 **end**
15 `Return` (SampleOutput)

Algorithm IV.5: SAMPLING ALGORITHM

Proof. The proof appears in the full version [26]. □

Proof of Theorem IV.6. Let notation be as in Theorem IV.6. (Termination): Let $\alpha = \frac{\epsilon - \delta}{\sqrt{N}}$. By definition of `SplitBox` there is an n such that the finitely many n-children of R in T_R have maximum side length at most α. Part (1) of Lemma IV.7 shows that the algorithm's breadth first search terminates at maximum depth n.
(Correctness): Let $\mathcal{M}$ be the set of boxes marked "stop". By part (2) of Lemma IV.7, $R = \cup_{M \in \mathcal{M}} M$. Let S be SAMPLEOUTPUT which was returned by the algorithm and Y be the set of center points of balls with form $B_{D_y - \delta}(y)$ in COVEREDREGIONS. By construction (algorithm line 8) any element $M \in \mathcal{M}$ has $M \subseteq B_\epsilon(s)$ for some $s \in S$ or $M \subseteq B_{D_y - \delta}(y)$ for some $y \in Y$. The properties of `MinDistance` guarantee that $X \cap (\cup_{y \in Y} B_{D_y - \delta}(y)) = \varnothing$. We have that $X \cap R \subseteq \cup_{s \in S} B_\epsilon(s)$. We also have $d_X(s) \leqslant \delta$ for all $s \in S$ by definition of `MinDistance`. Thus S is a (δ, ϵ)-sample of $X \cap R$. □

In practice, there are two quantities an optimal algorithm run should minimize: Calls to the relatively expensive `MinDistance` subroutine, and the number of points in the output sample. We can integrate geometric heuristics to reduce both quantities. These heuristics include:

- *Dynamic box splitting* - Adjust `SplitBox`(C) so that the largest intersection (by Lebesgue measure) of a box C with a region stored in COVEREDREGIONS is a box in `SplitBox`(C).
- *Dynamic sampling* - Refuse to add points to the output sample if their distance to the nearest point already in SAMPLEOUTPUT is less than some threshold.
- *Heuristic tree searching* - Place priority on first searching and applying `MinDistance` to the "largest" boxes at each level of depth in the search tree. A single run of

`MinDistance` has the potential to lead to a larger ball $B_{D_y-\delta}(y)$.

See [28] for an extended discussion of both the heuristics and implementation.

V. Examples

Algorithm IV.5 has been implemented and used to produce dense samples of varieties for further processing via PH. Example data is available at https://github.com/P-Edwards/sampling-varieties-data. Vietoris-Rips PH calculations were performed using the package Ripser [6] and persistence diagrams were produced using a plotting script in DIPHA [7].

In the following examples, regions of the persistence diagrams are highlighted according to Corollary II.13. Points in the highlighted region of an example's diagram correspond to homological features in the underlying variety, assuming the diagram was produced from a (δ, ϵ)-sample of a variety with homological feature size at least $2(\epsilon + \delta)$.

A. Clifford torus

The Clifford torus T is an embedding of the product of two circles, $S^1 \times S^1$, into $\mathbb{R}^4$. It is also a pure 2-dimensional algebraic variety defined by two equations in four variables:

$$T = \mathrm{V}_{\mathbb{R}}\left(x_1^2 + y_1^2 - \tfrac{1}{2}, x_2^2 + y_2^2 - \tfrac{1}{2}\right).$$

Since T is a torus, its Betti numbers are known theoretically to be $\beta_0 = 1, \beta_1 = 2$, and $\beta_2 = 1$. Note that T is compact as it is contained in the closed ball $\overline{B_1(0)}$ in $\mathbb{R}^4$. A sample of T was obtained by using Algorithm IV.5 to produce a $(10^{-7}, 0.14)$ sample of T (the bounding box used was $[-1, 1]^4$). The sample contains 5,689 points.

PH thresholded to a parameter value of 0.60 was subsequently calculated. The points in the persistence diagram represent features born before 0.60, and the points on the top edge represent features that do not die at 0.60 or earlier. The shaded region in Fig. 2 is derived from Corollary II.13. All points above and to the left of $(0.221, 0.56)$ are shaded.

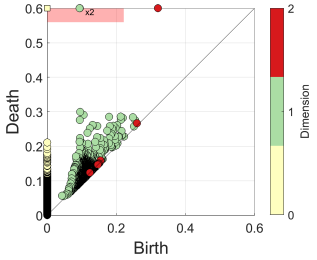

Fig. 2: PH results derived from sampling the Clifford torus. The sampling density is $(10^{-7}, 0.14)$. The estimated Betti numbers are $\beta_0 = 1$, $\beta_1 = 2$, and $\beta_2 = 1$.

B. Quartic surfaces

Restricting to the box $[-3, 3] \times [-3, 3] \times [-3, 3]$, we next consider the real algebraic varieties

$$V_1 = \mathrm{V}_{\mathbb{R}}\left(\begin{array}{c} 4x^4 + 7y^4 + 3z^4 - 3 - 8x^3 + 2x^2y - 4x^2 \\ -8xy^2 - 5xy + 8x - 6y^3 + 8y^2 + 4y \end{array}\right),$$

$$V_2 = \mathrm{V}_{\mathbb{R}}\left(\begin{array}{c} 144x^4 + 144y^4 - 225(x^2 + y^2)z^2 + 350x^2y^2 + 81z^4 \\ +x^3 + 7x^2y + 3x^2 + 3xy^2 - 4x - 5y^3 + 5y^2 + 5y \end{array}\right).$$

Both quartic equations define pure 2-dimensional varieties. Figure 3 displays visualizations of both V_1 and V_2 using the gathered samples allowing for a qualitative analysis. In particular, V_1 appears to be a sphere up to homotopy, with two distinct sphere-like features.

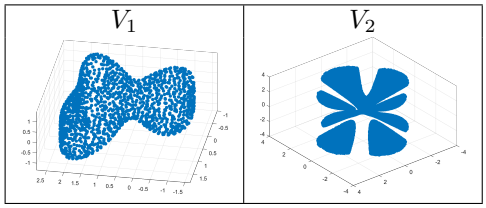

Fig. 3: Quartic surfaces sampled using Algorithm IV.5.

Samples produced for V_1 and V_2 contain 1,511 and 13,904 points respectively. The persistent homology results in Fig. 4(a) show that V_1 has homology features corresponding to a 2-sphere, with an additional 2-dimensional point which is relatively far away from the diagonal but not in the shaded region. The only homology features confirmed for V_2 in Fig. 4(b) are 5 connected components.

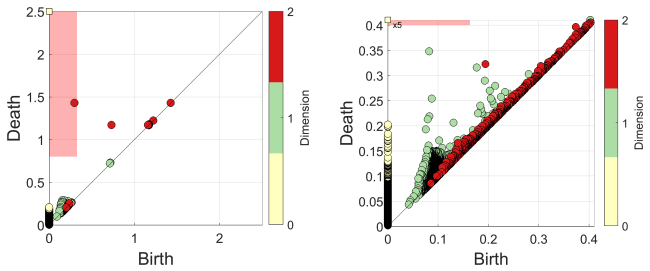

Fig. 4: (a) PH results for V_1 with sampling density $(10^{-7}, 0.20)$ and estimated Betti numbers $\beta_0 = 1$, $\beta_1 = 0$, $\beta_2 = 1$. (b) PH results for V_2 thresholded to a parameter value of 0.405 with sampling density $(10^{-7}, 0.10)$ and estimated Betti numbers $\beta_0 = 5$, $\beta_1 = 0$, $\beta_2 = 0$.

C. Deformable pentagonal linkages

Consider a regular pentagon in the plane consisting of links with unit length, with one of the links fixed to lie along the x-axis with leftmost point at $(0, 0)$. The space V_P of all possible configurations of this regular pentagon is a real algebraic variety. Farber and Schütz study this type of configuration space in [29], as well as providing an overview of its study. A specialization of their results shows that β_0 of V_P is 1, β_1 is 8, and β_2 is 1.

A description of the polynomials defining V_P is presented in [11, §6.2.2] where it is modelled as a compact pure 2-dimensional real algebraic variety in the six variables s_1, s_2, s_3 and c_1, c_2, c_3, namely:

$$V_P = V_{\mathbb{R}}\left(\begin{array}{ccc} s_1^2 + c_1^2 - 1, & s_2^2 + c_2^2 - 1, & s_3^2 + c_3^2 - 1, \\ (s_1 + s_2 + s_3)^2 + (1 + c_1 + c_2 + c_3)^2 - 1 & & \end{array}\right).$$

A $(10^{-7}, 1.12)$ sample of V_P was produced by first obtaining a $(10^{-7}, 1.0)$ sample using Algorithm IV.5. This sample

was then sub-sampled by iteratively choosing a point in the sample, removing all other points within 0.12 of the chosen point, and repeating this loop until all points in the subsample had no other points within distance 0.12. The sample contains 3,548 points. The PH results are summarized in Fig. 5. The points above and left of $(0.5, 1)$ in the diagram capture the theoretically expected homology for the configuration space.

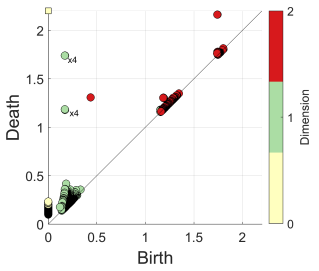

Fig. 5: Persistence diagram computed by sampling the configuration space of deformable pentagonal linkages. The sampling density is $(10^{-7}, 1.12)$.

VI. Conclusion and Future Work

The sampling algorithm presented in this paper is a first step towards systematizing the use of the TDA for obtaining geometric and topological information from algebraic varieties, including those that arise in applications. Our use of numerical algebraic geometry methods in producing dense samples is unique, and enables our algorithm to simultaneously satisfy both theoretical and practical constraints for applying TDA. The examples we provide in Section 5 illustrate how using the PH pipeline approach allows for the extraction of detailed information beyond Betti numbers on a real algebraic variety.

A step forward would be to derive and incorporate further information from the stratification structure of singular varieties into systematic TDA based analysis. Running the PH pipeline on individual strata after identifying them via stratification methods for samples (e.g. [9]) or algebraic methods (detailed in [36]) would result in an even more detailed summary of the variety. Another direction is to apply persistent homology of ellipsoids rather than ϵ-balls [13].

Our work also raises the natural question of computationally estimating a lower bound on the weak feature size of varieties. Future work will explore how to exploit the algebraic description for this purpose. Finally, it would be worthwhile to investigate the noise induced from sampling via homotopy continuation in the context of off-set varieties [38].

References

[1] N. Amenta and M. Bern. Surface reconstruction by voronoi filtering. *Discrete & Computational Geometry*, 22(4):481–504, 1999.

[2] N. Amenta, M. Bern, and M. Kamvysselis. A new voronoi-based surface reconstruction algorithm. In *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, pages 415–421. ACM, 1998.

[3] P. Aubry, F. Rouillier, and M. Safey El Din. Real solving for positive dimensional systems. *J. Symbolic Comput.*, 34(6):543–560, 2002.

[4] S. Basu. Computing the first few Betti numbers of semi-algebraic sets in single exponential time. *Journal of Symbolic Computation*, 41(10):1125–1154, 2006.

[5] D.J. Bates, J.D. Hauenstein, A.J. Sommese, and C.W. Wampler. *Numerically solving polynomial systems with Bertini*, volume 25. SIAM, 2013.

[6] U. Bauer. Ripser. https://github.com/Ripser/ripser, 2016.

[7] U. Bauer, M. Kerber, and J. Reininghaus. Dipha (a distributed persistent homology algorithm). Software available at https://github.com/DIPHA/dipha.

[8] U. Bauer, M. Kerber, and J. Reininghaus. Clear and compress: Computing persistent homology in chunks. In *Topological methods in data analysis and visualization III*, pages 103–117. Springer, 2014.

[9] P. Bendich, B. Wang, and S. Mukherjee. Local homology transfer and stratification learning. In *Proceedings of the twenty-third annual ACM-SIAM symposium on Discrete Algorithms*, pages 1355–1370. Society for Industrial and Applied Mathematics, 2012.

[10] M. Berger, A. Tagliasacchi, L. Seversky, P. Alliez, J. Levine, A. Sharf, and C. Silva. State of the art in surface reconstruction from point clouds. In *EUROGRAPHICS star reports*, volume 1, pages 161–185, 2014.

[11] D.A. Brake, D.J. Bates, W. Hao, J.D. Hauenstein, A.J. Sommese, and C.W. Wampler. Algorithm 976: Bertini_real: Numerical decomposition of real algebraic curves and surfaces. *ACM Trans. Math. Softw.*, 44(1):10, 2017.

[12] P. Breiding and O. Marigliano. Sampling from the uniform distribution on an algebraic manifold. *arXiv:1810.06271*, 2018.

[13] P. Breiding, S. Kališnik, B. Sturmfels, and M. Weinstein. Learning algebraic varieties from samples. *Revista Matematica Complutense*, 31(3):545–593, 2018.

[14] P. Bubenik, V. De Silva, and J. Scott. Metrics for generalized persistence modules. *Foundations of Computational Mathematics*, 15(6):1501–1531, 2015.

[15] P. Bubenik and J.A. Scott. Categorification of persistent homology. *Disc. & Comput. Geom.*, 51(3):600–627, 2014.

[16] G. Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308, 2009.

[17] F. Chazal, D. Cohen-Steiner, and A. Lieutier. A sampling theory for compact sets in euclidean space. *Discrete & Computational Geometry*, 41(3):461–479, 2009.

[18] F. Chazal and A. Lieutier. Weak feature size and persistent homology: computing homology of solids in $\mathbb{R}^n$ from noisy data samples. In *Proceedings of the twenty-first annual symposium on Computational geometry*, pages 255–262. ACM, 2005.

[19] F. Chazal and B. Michel. An introduction to topological data analysis: fundamental and practical aspects for data scientists. *arXiv:1710.04019*, 2017.

[20] C. Chen and M. Kerber. Persistent homology computation with a twist. In *Proceedings 27th European Workshop on Computational Geometry*, volume 11, 2011.

[21] D. Cohen-Steiner, H. Edelsbrunner, and J. Harer. Stability of persistence diagrams. *Discrete & Computational Geometry*, 37(1):103–120, 2007.

[22] F. Cucker, T. Krick, and M. Shub. Computing the homology of real projective sets. *Foundations of Computational Mathematics*, pages 1–42, 2016.

[23] V. De Silva and R. Ghrist. Coverage in sensor networks via persistent homology. *Algebraic & Geometric Topology*, 7(1):339–358, 2007.

[24] T. Krishna Dey, X. Ge, Q. Que, I. Safa, L. Wang, and Y. Wang. Feature-preserving reconstruction of singular surfaces. In *Computer Graphics Forum*, volume 31, pages 1787–1796. Wiley Online Library, 2012.

[25] J. Draisma, E. Horobeţ, G. Ottaviani, B. Sturmfels, and R. Thomas. The Euclidean distance degree. In *SNC 2014—Proceedings of the 2014 Symposium on Symbolic-Numeric Computation*, pages 9–16. ACM, New York, 2014.

[26] E. Dufresne, P.B. Edwards, H.A. Harrington, and J.D. Hauenstein. Sampling real algebraic varieties for topological data analysis. `arXiv:1802.07716`, 2018.

[27] H. Edelsbrunner and J. Harer. *Computational topology: an introduction*. AMS, 2010.

[28] P.B. Edwards. Topological data analysis for real algebraic varieties. Master's thesis, University of Oxford, 2016.

[29] M. Farber and D. Schütz. Homology of planar polygon spaces. *Geometriae Dedicata*, 125(1):75–92, 2007.

[30] D. Freedman. An incremental algorithm for reconstruction of surfaces of arbitrary codimension. *Computational Geometry*, 36(2):106–116, 2007.

[31] J.H.G. Fu. Tubular neighborhoods in Euclidean spaces. *Duke Mathematical Journal*, 52(4):1025–1046, 1985.

[32] R. Ghrist. Barcodes: the persistent topology of data. *Bulletin of the American Math. Society*, 45(1):61–75, 2008.

[33] E. Gross, H.A. Harrington, Z. Rosen, and B. Sturmfels. Algebraic Systems Biology: A Case Study for the Wnt Pathway. *Bulletin of Mathematical Biology*, 78(1):21–51, 2016.

[34] O. Hanner. Some theorems on absolute neighborhood retracts. *Ark. Mat.*, 1:389–408, 1951.

[35] A. Hatcher. *Algebraic Topology*. Cambridge University Press, 2002.

[36] J.D. Hauenstein and C.W. Wampler. Isosingular sets and deflation. *Foundations of Computational Mathematics*, 13(3):371–403, 2013.

[37] J.D. Hauenstein. Numerically computing real points on algebraic sets. *Acta Appl. Math.*, 125:105–119, 2013.

[38] E. Horobet and M. Weinstein. Offset hypersurfaces and persistent homology of algebraic varieties. *arXiv:1803.07281*, 2018.

[39] F. Le Gall. Powers of tensors and fast matrix multiplication. In *ISSAC 2014—Proceedings of the 39th International Symposium on Symbolic and Algebraic Computation*, pages 296–303. ACM, New York, 2014.

[40] S. Lojasiewicz. Triangulation of semi-analytic sets. *Ann. Scuola Norm. Sup. Pisa Cl. Sci. (3)*, 18:449–474, 1964.

[41] S. Martin, A. Thompson, E.A. Coutsias, and J.-P. Watson. Topology of cyclo-octane energy landscape. *The journal of chemical physics*, 132(23):234115, 2010.

[42] S. Martin and J.-P. Watson. Non-manifold surface reconstruction from high-dimensional point cloud data. *Computational Geometry*, 44(8):427–441, 2011.

[43] R. Mendoza-Smith and J. Tanner. Parallel multi-scale reduction of persistent homology filtrations, 2017.

[44] N. Milosavljević, D. Morozov, and P. Skraba. Zigzag persistent homology in matrix multiplication time. In *Proceedings of the twenty-seventh annual symposium on computational geometry*, pp. 216–225, ACM, 2011.

[45] A.P. Morgan and A.J. Sommese. Coefficient-parameter polynomial continuation. *Appl. Math. Comput.*, 29(2, part II):123–160, 1989.

[46] B. Mourrain and J.P. Pavone. Subdivision methods for solving polynomial equations. *Journal of Symbolic Computation*, 44(3):292–306, 2009.

[47] M. Mustaţă. Graded betti numbers of general finite subsets of points on projective varieties. *Le Matematiche*, 53(3):53–81, 1998.

[48] P. Niyogi, S. Smale, and S. Weinberger. Finding the Homology of Submanifolds with High Confidence from Random Samples. *Discrete & Computational Geometry*, 39(1-3):419–441, 2008.

[49] N. Otter, M.A. Porter, U. Tillmann, P. Grindrod, and H.A. Harrington. A roadmap for the computation of persistent homology. *EPJ Data Science*, 6(1):17, 2017.

[50] S.Y. Oudot. *Persistence theory: from quiver representations to data analysis*, volume 209. American Mathematical Society, 2015.

[51] F. Rouillier, M.-F. Roy, and M. Safey El Din. Finding at least one point in each connected component of a real algebraic set defined by a single equation. *J. Complexity*, 16(4):716–750, 2000.

[52] P. Scheiblechner. On the complexity of deciding connectedness and computing betti numbers of a complex algebraic variety. *Journal of Complexity*, 23(3):359–379, 2007.

[53] A. Seidenberg. A new decision method for elementary algebra. *Ann. of Math. (2)*, 60:365–374, 1954.

[54] E.C. Sherbrooke and N.M. Patrikalakis. Computation of the solutions of nonlinear polynomial systems. *Computer Aided Geometric Design*, 10(5):379–405, 1993.

[55] A. Sommese and C. Wampler. *The Numerical solution of systems of polynomials arising in engineering and science*, volume 99. World Scientific, 2005.

[56] A. Zomorodian and G. Carlsson. Computing persistent homology. *Disc. & Comput. Geom.*, 33(2):249–274, 2005.